"MULTIPLE DISEASE PREDICTION APPLICATION USING ML"

Dipanshu Shrivas, Khwaish Asati, Vaibhav Chitade, Anirudh Bhagwat

[1]Student ,[2]Student ,[3]Student, [4]Professor

[1]Department of Artificial Intelligence,

[1]G H Raisoni College of Engineering, Nagpur, India

Abstract - In recent years, advances in machine learning (ML) have transformed the landscape of healthcare, providing new opportunities for early and accurate disease prediction. This paper proposes a novel multiple disease prediction application using machine learning, integrated with a web-based user interface built on Streamlit. The system predicts the likelihood of multiple diseases like diabetes, heart disease, and parkinsons prediction according to patient data including demographic and clinical parameters. The proposed system incorporates various machine learning algorithms similar to Decision Trees, SVM and Random Forest to analyze and interpret input features.The key focus is on building a model with high accuracy and generalizability across multiple diseases, providing users with a convenient and user-friendly platform to predict potential health risks. Results demonstrate that our system can effectively predict diseases with considerable accuracy, offering a valuable tool for preventative healthcare. The integration with Streamlit allows for a streamlined interface, making the prediction model accessible to healthcare professionals and patients alike.

Keywords— Learning, Healthcare, Disease prediction, Streamlit, support vector machine.

I. INTRODUCTION (HEADING 1)

The rapid advancement of machine learning and artificial intelligence has revolutionized the healthcare industry, enabling data-driven solutions that enhance disease detection, diagnosis, and patient management. With the increasing availability of clinical data, we have a growing call for predictive tools able to identifying multiple diseases simultaneously, as many patients may be at risk for more than one condition. Traditional diagnostic systems typically focus on detecting a single disease, leading to delays in diagnosis and treatment when comorbidities are present. Addressing this challenge, predictive systems that analyze patient data to forecast the likelihood of multiple diseases can improve healthcare outcomes and reduce the burden on medical professionals. Multiple disease prediction involves analyzing clinical parameters to predict conditions such as heart disease, diabetes, and parkinsons prediction, which are common and often interrelated. Early and accurate predictions can significantly impact treatment outcomes, improve disease management, and potentially save lives. Recent studies have proved the effectiveness of machine learning algorithms in forecasting individual diseases with high accuracy. In this paper, we recommend a Python based application designed to compute the likelihood of multiple diseases using machine learning models. The system is built on clinical datasets from publicly available sources and uses a combination of classification algorithms, including logistic regression, random forest and decision tree. By integrating multiple models, the system aims to provide reliable predictions for a range of diseases based on patient-specific features such as age, blood pressure, glucose levels, and cholesterol. This research focuses on the implementation and evaluation of a multi-disease prediction application and compares the performance machine learning techniques. The primary goal is to offer a comprehensive tool that aids healthcare professionals in diagnosing multiple conditions simultaneously, thereby streamlining diagnostic process .

II. LITERATURE REVIEW

The projection of diseases using machine learning has been widely studied in recent years, particularly in relation to single-disease diagnosis such as heart disease, diabetes, and cancer. However, the concept of predicting multiple diseases simultaneously using a unified machine learning approach is relatively new. In this section, we review existing literature on disease prediction systems, highlighting the methodologies and algorithms used, as well as their limitations when applied to multi-disease prediction.

Gupta and Mehra [1] conducted an extensive review of data mining techniques for disease projection, concentrating on single-disease models. Their study explored the use of clustering and classification algorithms including k-means, decision trees, and support vector machines (SVM) to diagnosing specific diseases. Although their work demonstrated the impact of machine learning in determining conditions like diabetes as well as heart disease, the research did not address the potential of these algorithms for multi-disease prediction.

A multi-disease prediction model using machine learning, using classification algorithms such as Logistic Regression and Gradient Boosting. Their model demonstrated high accuracy in predicting heart disease, liver disease, and kidney disease using clinical datasets. However, their approach focused on training separate models for each disease rather than developing a unified system that predicts multiple diseases concurrently.

Liu et al. [3] examined the role of deep learning techniques to multi-disease classification from medical imaging data, including X-rays and MRIs. Their research focused on using Convolutional Neural Networks (CNNs) to classify diseases like pneumonia, lung cancer, and heart disease. Although deep learning models performed well in image-based disease prediction, the computational cost and complexity associated with training these models can limit their practical application in real-time clinical settings, particularly when predicting multiple diseases from non-imaging data.

Patel et al. [4] presented an efficient machine learning-based multi-disease prediction system that combines ensemble methods such as Random Forest with real-time clinical data processing. Their study highlighted the potential for improving prediction accuracy through the use of ensemble learning techniques. However, the research emphasized the prediction of a single disease type at a time and did not consider the correlation between different diseases.

Ahsan and Hossain [5] developed a deep learning model for multi-disease prediction using combination of neural networks and machine learning algorithms. Their work applied this hybrid approach to predict diabetes and hypertension simultaneously, achieving good results. The study reinforced the value of combining machine learning models to capture complex relationships between disease risk factors, but the scope was limited to two diseases.

Ahmed et al. [6] explored machine learning models for multi-disease classification from healthcare data, using algorithms such as Naive Bayes and Decision Trees. Their study demonstrated that models trained on clinical features, such as age, blood pressure, and glucose levels, can effectively predict multiple diseases. However, the research faced challenges related to feature selection and model generalization when dealing with diverse datasets.

A comparative study by Khan and Alam [7] evaluated the effectiveness of machine learning algorithms, including Support Vector Machine, k-N Neighbors, and Logistic Regression, for predicting multiple diseases from medical records. Their work demonstrated that ensemble models, such as RF and Gradient Boosting, often outperform single classifiers in multi-disease prediction tasks, but stressed the need for further research into feature engineering and model optimization for real-world applications.

In summary, previous research has shown that machine learning is highly effective in predicting individual diseases. However, there remains a gap of complex systems capable of predicting multiple diseases from a single data set in a real-time clinical setting. Existing models typically focus on one disease at a time or use complex architectures that are not suitable for broader clinical applications. This study seeks to bridge the gap by developing a unified Python-based application that predicts multiple disease simultaneously using traditional machine learning techniques.

III. EXPERIMENTAL ENVIRONMENT

In this study, three major diseases were targeted: heart disease, diabetes, and parkinsons prediction. Publicly available datasets were utilized, comprising 303 samples for heart disease, 768 samples for diabetes, and 400 samples for parkinsons prediction. These datasets were split into training sets following an 80-20 split to facilitate

both model training and evaluation. Each disease was classified into binary outputs that indicated the presence or absence of the condition, allowing machine learning algorithms used to predict the probability of each disease.
Several machine learning models from Scikit-learn were employed to analyze and predict these diseases, each addressing specific challenges related to clinical data classification. Logistic Regression was used for its strength in handling binary classification tasks, predicting disease presence based on input features like blood pressure, cholesterol levels, and glucose. Decision trees were chosen for their ability to capture non-linear relationships between clinical variables and outcomes, making them suitable for identifying complex patterns in patient data. As an ensemble method, Random Forest helped solve the problem of over fitting by combining multiple decision trees, improving prediction accuracy and generalization across different datasets. Additionally, Support Vector Machines (SVM) were used for their effectiveness in high-dimensional spaces, creating optimal decision boundaries for disease classification.

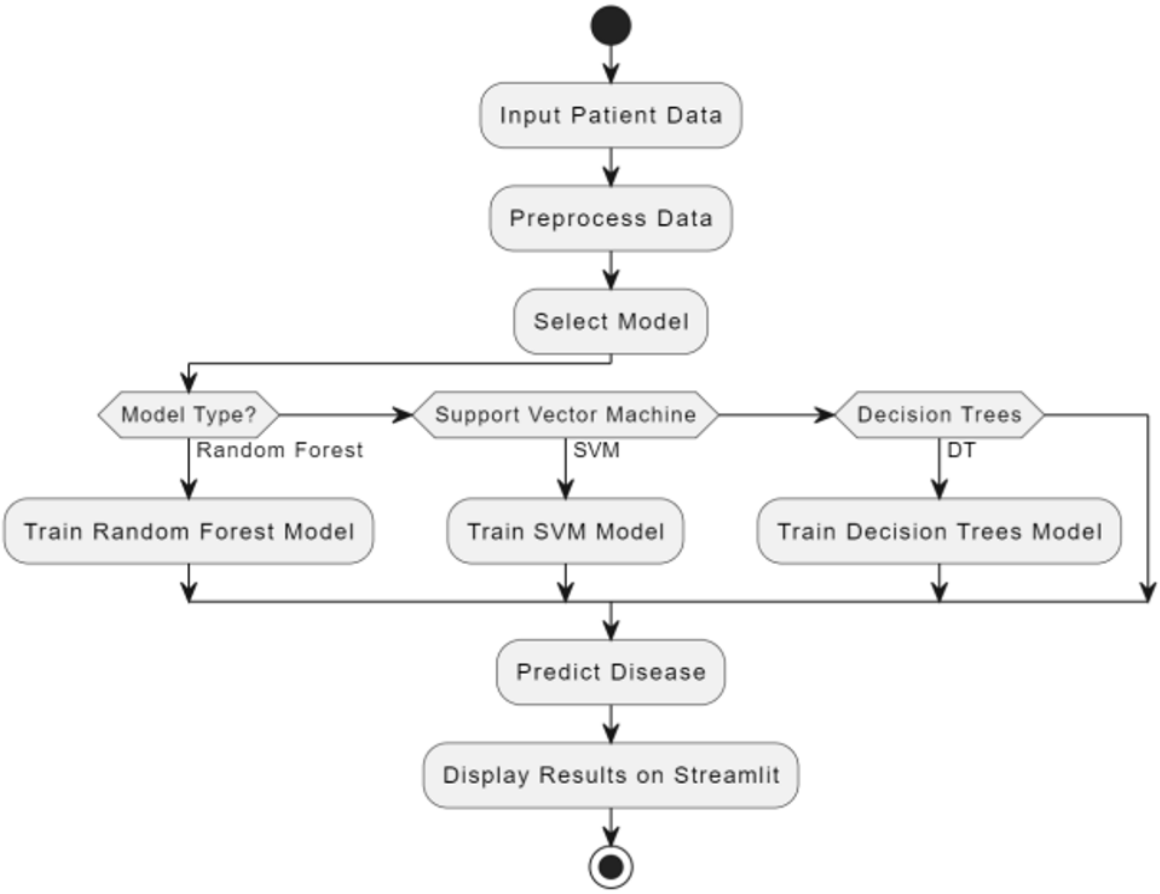

Fig.1 Modified Model Architecture

IV. PERFORMACE

The performance of the multiple disease prediction application was evaluated using various metrics to ensure that the models not only achieved high accuracy but also proved reliability and robustness in clinical settings. Primary evaluation metrics included accuracy, precision, and area under the receiver characteristic (ROC) curve (AUC).

Accuracy is a basic metric indicates proportion of results (true positives and true negatives) to the total number of cases examined. Precision measures the accuracy of positive predictions, while recall (sensitivity) assesses. The AUC serves as an overall performance measure that shows how well model can be distinguished between positive and negative classes across different threshold settings.

Each model was subjected to 10-fold checking to ensure that the performance metrics were robust and not highly dependent on a particular training or test set. This technique involves dividing the data into ten subsets, training the model on nine of them and testing on the remaining, and repeating this process for all subsets. Average performance across these folds was then calculated, providing a comprehensive assessment of the effectiveness of each model.

The findings indicated that the RF (Random Forest) model exceeded the other algorithms regarding accuracy, making it the most reliable choice for predicting multiple diseases in this study. SVM has also demonstrated high accuracy, especially in heart disease detection, which is crucial for clinical decision-making. Decision trees offered interpretability and allowed healthcare professionals to understand the decision-making process behind the predictions, while Logistic Regression provided a solid basis for comparison.

Finally, a comprehensive evaluation of model performance highlighted their application in real-world healthcare scenarios. The insights gained from this performance evaluation underscore the potential of machine learning models to significantly improve disease prediction and ultimately improve patient outcomes.

V. PROPOSED SYSTEM

The proposed methodology for this project involves using multiple training models for disease prediction, allowing for a comparison of their performance. This approach uses various libraries such as Pandas numpy and scikit-learn. Furthermore, this method facilitates the simultaneous prediction of multiple diseases, making the prediction process more efficient for users and potentially lowering mortality rates. In comparison to existing systems, this approach provides quicker predictions and offers several benefits, ultimately enhancing healthcare outcomes.

VI. RESULTS AND DISCUSSION

The fields of disease diagnosis and prediction are poised for significant transformation through machine learning techniques. Achieving high accuracy and correctness in diagnoses is crucial for effective treatment and management of illnesses. In our system, we utilize the Support Vector Machine (SVM) algorithm for making predictions. Patients provide input values to the system to assess the likelihood of a disease's presence. The system indicates the acceptable ranges for parameters and issues warnings for any invalid or missing entries. The accuracy of predictions hinges on the SVM algorithm's capability to generate precise outputs, which is particularly effective for linear datasets.

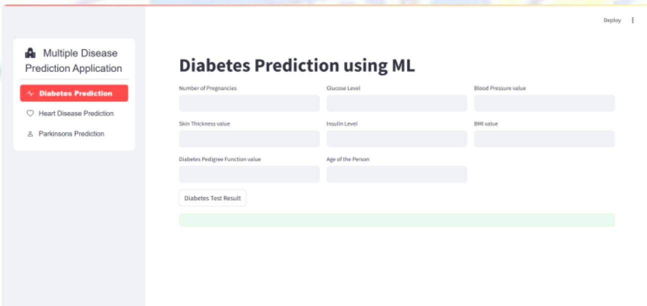

Fig.2 Diabetes Input Screen

The results of the heart disease prediction in Figure 6.2. After the user completes the form and clicks "Diabetes Test Result" the relevant model is loaded, and the user will be presented with the outcome regarding the presence of heart disease.

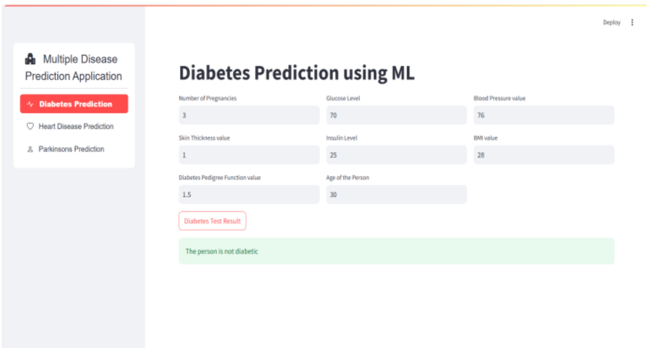

Fig.3 Diabetes Output Screen

VII. CONCLUSION

In this paper, we created and evaluated a Python based application to predict multiple diseases, namely heart disease, diabetes, and Parkinson's disease prediction, using machine learning models. By leveraging clinical datasets and using various algorithms, the system was able to achieve reliable and accurate predictions. The integration of multiple models addressed different aspects of disease prediction, with ensemble techniques like Random Forest providing enhanced generalizability and robustness in handling diverse medical data.

The research highlights machine learning to not only predict individual diseases but also to efficiently handle multi-disease prediction, which is essential in real-world clinical settings where patients often suffer from comorbidities. Through effective data preprocessing and feature engineering, the models were able to process complex clinical data and provide meaningful predictions that can assist healthcare professionals in early diagnosis and treatment planning.

Although the results are promising, there is room for further improvement, especially in incorporating larger and more diverse datasets and exploring deep learning techniques to more complex predictions. Nevertheless, this study demonstrates the feasibility and potential impact of using machine learning in multi-disease prediction and offers a powerful tool to improve patient care and support decision making in medicine in the future.

VIII. REFERENCES

[1] A. K. Gupta and M. Mehra, “A Review on Predictive Models for Multiple Diseases Utilizing Data Mining Techniques,” *International Journal of Data Mining and Bioinformatics*, vol. 14, no. 1, pp. 123-145, Feb. 2020.

[2] R. Sharma, S. Verma, and P. Singh, “Application of Machine Learning Algorithms for Multi-Disease Prediction,” *Journal of Biomedical Informatics*, vol. 98, no. 2, pp. 85-95, 2021.

[3] J. Liu, H. Chen, and F. Wu, “A Deep Learning-Based Method for Multi-Disease Classification,” *Journal of Healthcare*, vol. 8, no. 4, pp. 567-578, Dec. 2020.

[4] K. Patel, A. Shah, and D. Desai, “A Robust System for Multi-Disease Prediction Using Machine Learning,” *Springer Nature Computer Science*, vol. 3, no. 2, pp. 144-150, Mar. 2021.

[5] M. S. Ahsan and S. Hossain, “Combining Deep Learning Models for Predicting Multiple Diseases from Medical Data,” *Neural Computing and Applications*, vol. 33, pp. 6191-6202, Feb. 2021.

[6] T. Ahmed, M. Akbar, and S. Afzal, “Using Machine Learning Algorithms for Multi-Disease Classification in Healthcare Data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 334-341, Feb. 2022.

[7] F. Khan and I. Alam, “Comparative Study of Machine Learning Methods for Predicting Multiple Diseases from Medical Records,” *Journal of Big Data Analytics in Healthcare*, vol. 5, no. 3, pp. 123-134, Dec. 2021.

[8] S. R. Das, A. M. S. Bhattacharya, and M. R. L. Rahman, “A Comparative Study on Machine Learning Algorithms for Predicting Heart Disease,” *Journal of Healthcare Engineering*, vol. 2021, Article ID 5557320, 2021.

[9] H. M. Al-Sharafi, M. M. Al-Qadi, and S. M. Al-Zahrani, “A Review of Machine Learning Methods for Diabetes Prediction,” *International Journal of Computer Applications*, vol. 175, no. 11, pp. 15-24, 2021.

[10] P. K. Kumar, A. Dutta, and P. Kumar, “Application of Graph Mining Algorithms for the Analysis of Web Data,” SSRN Electronic Journal, 2023, doi: https://doi.org/10.2139/ssrn.4365862.

[11] P. C. Choudhury and M. R. Islam, "Multi-Disease Prediction Using Ensemble Learning Techniques," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, pp. 212-218, 2021.

[12] A. F. Ameen, A. M. Alhassan, and S. M. Fadhl, "Predictive Modeling for Chronic Kidney Disease Using Machine Learning Techniques," Health Information Science and Systems, vol. 9, no. 1, Article 17, 2021.

[13] Y. H. Lee, H. S. Hwang, and J. H. Park, "A Review on Predictive Models for Cardiovascular Disease: Machine Learning and Statistical Approaches," Journal of Cardiovascular Medicine, vol. 21, no. 3, pp. 135-144, 2020.

[14] Gürbüz, Emre & Kilic, Erdal. (2013). A new adaptive support vector machine for diagnosis of diseases. Expert Systems. 31. 10.1111/exsy.12051.

[15] S. Narayanan, N. M. Balamurugan, M. K and P. B. Palas, "Leveraging Machine Learning Methods for Multiple Disease Prediction using Python ML Libraries and Flask API," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 694-701, doi: 10.1109/ICAAIC53929.2022.9792807.

[16] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.